# Metacognition for Well-Learned Information

Talira Kucina

A report submitted as a partial requirement for the degree of Bachelor of Psychological Science with Honours at the University of Tasmania, 2017

**Statement of Sources**

I declare that this report is my own original work and that contributions of others have been duly acknowledged.

Signed: _____

Date: 19 October 2017

**Acknowledgements**

First and foremost, thank you to my supervisor Dr Matthew Palmer for sharing your knowledge and insight, for supporting and challenging me, and for your tireless work and continual guidance. Thank you for encouraging and motivating me with your passionate approach to research. It has been an absolute privilege to work with you on this project.

I would like to thank Dr Jim Sauer for advice regarding this study and Simon Bury at Flinders University for assisting with data collection.

A special thank you to Matthew Gretton for creating a program with which to run my experiment. I am incredibly grateful for your time and attention to detail.

Thank you to Laura Brumby and Glenys Holt for the support, advice, and encouragement you have given throughout the year. Laura, I truly appreciate your help in programming part of this experiment.

I would like to thank the psychology staff in Launceston for the support and encouragement shown during the year.

I am very grateful to every participant who offered their time to take part in this study – thank you for making this project possible.

Thank you to my fellow honour students for making this year so enjoyable, with many wonderful memories.

Finally, I would like to thank my family – my mother and father, and sister, Sharissa – for believing in me and for being so incredibly supportive not only this year, but those preceding it.

## Table of Contents

## List of Tables and Figures

Metacognition for Well-Learned Information

Talira Kucina

Word Count: 9906

## Abstract

The current study investigated the effects of criterion learning and lag on metacognitive accuracy and memory performance. Forty-six participants (27 female) aged between 18-63 years studied lists of Lithuanian-English word pairs presented in either groups of nine (short lag) or 36 items (long lag). Participants engaged in practice testing for the items until they were correctly recalled one, three, or nine times. Participants made judgments of learning regarding the likelihood they would remember the item in about 7 days, where they would engage in a cued recall task. Judgements were recorded on either binary or continuous scales. Absolute accuracy was greater for the long lag compared to short lag ($p < .001$), which led to substantial overconfidence in the short lag ($p < .001$). Relative accuracy was superior in the long lag compared to the short lag ($p < .001$). Final cued recall performance was higher in the long lag and as criterion increased. These findings suggest students' memory performance will benefit from the repeated successful recall of information. Crucially, if students employ a longer lag, not only will their memory be enhanced but they will likely display superior metacognitive monitoring of their learning.

Metacognition is important in many facets of daily life and involves the understanding, monitoring, and control of one's own cognitive processes (Flavell, 1979). Other aspects include metacognitive knowledge and regulatory capacity (Veenman, Van Hout-Wolters, & Afflerbach, 2006). Judgments of learning (JOLs) are a type of metacognitive judgment requiring the individual to predict the likelihood that they will correctly recall a given item or information at a future time, and are therefore considered a way to monitor learning (Frank & Kuhlmann, 2017). As an example, consider someone thinking about the items they have on a shopping list; there are several items needed and thus they mentally go through the list. The individual may consider whether they will be able to recall each item when in the supermarket, without writing the items down. If the person thinks they will remember every item and once in the supermarket they do, then their metacognitive capacity is considered good. However, metacognition is not just about remembering information, it also involves awareness of when one is unlikely to recall information and what can be done about this. Therefore, if the shopper thinks they will have difficulty recalling all the items they need (as this has been the case in the past), they may generate a strategy to help. They may simply write the list down, or if they were highly determined to remember the items, without such a list, they could engage in a strategy such as self-testing. This not only enhances memory, but allows the individual to establish which items they do and do not remember, meaning they can subsequently pay more attention to these harder to recall items (Kornell & Son, 2009).

The judgments alluded to above are not formed on the basis of knowledge alone, but include the experiences and feelings of the individual (Finn & Tauber, 2015). For the majority of individuals, metacognitive knowledge and abilities

develop to some degree as a result of their interactions with others, particularly those with teachers and through schooling (Veenman et al., 2006). There is however much variability in metacognitive awareness between individuals and across various contexts. Ultimately, metacognitive accuracy is contingent upon the correct prediction of items that will and will not be recalled at a later timepoint.

It is imperative to understand and increase the accuracy of JOLs as they can influence the degree of revision engagement, study time allocation, and ultimately performance (Metcalfe & Finn, 2008; Van Overschelde & Nelson, 2006). Consequently, inaccurate JOLs may result in the ineffective utilisation of study time and can impede effective learning. Crucially, both the study techniques used and the scale for JOL measurement can impact different types of metacognitive accuracy including absolute accuracy and relative accuracy. Both types of accuracy are important, for example if relative accuracy is poor then people are less likely to prioritise the material which would benefit most from subsequent attention, and when poor absolute accuracy results in overconfidence, people are unlikely to engage in adequate study for sufficient learning to occur (Kornell & Bjork, 2008).

The current experiment examined the effects of study strategies, such as criterion learning and lag (i.e., spaced study), on memory performance and metacognitive accuracy. It also investigated whether metacognitive accuracy would be influenced by the type of scale JOLs were measured on. That is, whether greater metacognitive accuracy would be found for dichotomous compared to continuous measurement scales for information that is well-learned.

## Theoretical Frameworks for JOL Assignment

Many contemporary theories of metacognition posit JOLs are inferential in nature (Koriat & Ma'ayan, 2005; Pyc, Rawson, & Aschenbrenner, 2014), and as such

have been guided by the cue-utilisation paradigm (Koriat, 1997). JOLs are based on cues pertaining to the current task and are used to infer the state of one's memory rather than using the strength of memory itself (Pyc & Rawson, 2012). Accordingly, JOLs may be shaped by the three cue types put forward by Koriat (1997); intrinsic, extrinsic, and mnemonic. Firstly, intrinsic cues relate to features of the information or material itself, for example item difficulty (Thomas, Finn, & Jacoby, 2016). Extrinsic cues include both conditions of learning and encoding, while mnemonic cues have an indirect influence on JOLs and include factors such as cue familiarity and ease of processing (Koriat, 1997).

The *anchoring hypothesis* (Scheck & Nelson, 2005) suggests the JOL values people assign during the initial stages of learning new information partly form the basis of future JOLs given on 0-100% probability scales. JOLs are frequently measured on this type of scale, where 0 indicates complete certainty the item will not be remembered and 100 reflects complete certainty it will be. Various initial anchor points have been proposed, starting as low as 20-30% (Serra & England, 2012) with many others favouring values around 50% (Connor, Dunlosky, & Hertzog, 1997; Dunlosky, Serra, Matvey, & Rawson, 2005; Hanczakowski et al., 2013) or ranging somewhere between these values (30-50%; Rast & Zimprich, 2009). While some claim the anchor is purely an arbitrary value on which JOLs are based (Zawadzka & Higham, 2016), others claim it may be the mechanism accounting for inaccurate metacognitive judgments, for example, when JOLs underestimate performance over multiple practice or study trials, which is termed the *underconfidence-with-practice (UWP)* effect (Scheck, Meeter, & Nelson, 2004). According to this idea, the UWP effect is due to insufficient adjustment from the initial anchor point as a result of true cognitive bias and inadequate knowledge of what or how much one knows.

Regardless of the basis for the anchor value, it still acts as a cue informing later JOLs. However, it is important to note that there are certain conditions under which the anchor point may not be influential (Hanczakowski et al., 2013; Zawadzka & Higham, 2015), as later discussed.

## Metacognitive Measures

As aforementioned there are two types of accuracy regarding metacognitive monitoring; the first is absolute accuracy, or calibration, which represents the degree to which perception of performance corresponds to actual performance (Keren, 1991). Therefore, a probabilistic interpretation of perfect calibration entails, for instance, that of all the items given a 60% rating, 60% of these will subsequently be recalled (Van Overschelde & Nelson, 2006). The second is relative accuracy, or resolution, which refers to the ability to accurately differentiate between items that will and will not be remembered at some future time (Yaniv, Yates, & Smith, 1991). Resolution is sound when one accurately determines the information they are likely to recall relative to that which they will not recall (Ariel & Dunlosky, 2011). It should be noted that one may be well calibrated, but have poor resolution, or alternatively they may have good resolution yet be poorly calibrated.

## Influence of Study Strategies on Memory and JOLs

Study strategy effectiveness has been extensively researched with evidence that practice testing, where material is repeatedly tested or recalled rather than merely restudied, is highly beneficial for learning (Cull, 2000; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). A similar yet clearly distinguishable variant of practice testing is the concept of criterion learning. This is defined as the number of times each item or piece of information must be successfully recalled in a study session (Vaughn & Rawson, 2011). In practical terms this distinction means that

attempting to recall information six times, not necessarily successfully, constitutes retrieval practice, while criterion learning requires the attempts to be successful six times, even if this means the total number of attempts is 10, 12, or any other number greater than six. An example is attempting to recall the definitions of a list of words and stopping once each word has received six attempted recalls (i.e., retrieval practice) contrasted with continuing to attempt the definitions until all have been successfully recalled six times (i.e., criterion learning). When experimentally imposed, higher criterion levels enhance memory, such as one study where researchers had participants correctly recall information up to eight times (Karpicke & Roediger, 2007) and another where five was the maximum criterion (Vaughn & Rawson, 2011). Additionally, there is evidence that students understand and voluntarily use criterion learning; in one survey approximately 65% of respondents reported that they continue studying flashcards until they correctly remember the content at least once (Wissman, Rawson, & Pyc, 2012). In sum, attempting to retrieve an item more times compared to fewer times enhances performance on tasks reliant upon memory, and possibly even more so when the retrieval attempt is successful (Pyc & Rawson, 2009).

Regarding metacognitive monitoring during the repeated presentation of material, some researchers have suggested people maintain some understanding of the benefits of retrieval practice (Kornell & Bjork, 2007; Pyc & Rawson, 2012). However, there is also evidence that people have difficulty scaling strong memories (Mickes, Hwe, Wais, & Wixted, 2011) such as when repeated retrieval occurs. It is often found that accurate knowledge concerning the memorial benefits of repeated retrieval is either poor or inadequately applied to repeated learning trials. This is commonly evidenced by the UWP effect whereby people underestimate their

learning relative to performance over practice trials (Karpicke, 2009; Koriat, Sheffer, & Ma'ayan, 2002; Pyc et al., 2014). That is, although JOLs often increase over trials, it tends to be to an inaccurate degree. Hence, people may be aware that learning to a higher criterion enhances memory, but unaware of the magnitude of this benefit. The limited research focusing specifically on criterion learning and metacognition has produced various outcomes. Some suggest criterion learning does not necessarily enhance overall metacognitive accuracy relative to pure restudy (Karpicke, 2009) and others finding that higher criterion were associated with slightly superior metacognitive accuracy (Pyc et al., 2014). Finally, to clarify, the above does not necessarily refer to underconfidence only, but may be manifest as overconfidence. An example of such overconfidence is when people fail to recognise that learning to increasingly higher criterion levels provides diminishing returns in terms of memory performance (Pyc & Rawson, 2012).

Another effective study technique is distributed practice where the study of material is temporally spaced, either during a single study period or over multiple sessions on multiple different days (Dunlosky et al., 2013; Kornell & Bjork, 2007). Of particular importance in the current experiment was lag, which is specifically defined as the number of items between one presentation of a certain item and the next presentation of that item (Logan, Castel, Haber, & Viehman, 2012). An example demonstrating a shorter lag is learning concepts via a smaller group of flashcards comparative to a longer lag which would entail using a larger group of flashcards. Thus, for the short lag, one may divide the larger group into smaller groups, as such a total of 45 cards may be split into five groups of nine cards each. The individual would begin by repeatedly studying a single group before moving onto the next group. In the longer lag, all cards would be studied in one group (i.e., a single group

of 45 cards). Several studies (Cull, 2000; Pavlik & Anderson, 2005; Pyc & Rawson, 2009) have established the memorial benefits of a longer lag relative to a shorter lag. Of note, infinitely longer lag times will not necessarily lead to superior memory performance (Küpper-Tetzel & Erdfelder, 2012). Eventually, a lag becomes too long and this is detrimental to subsequent memory performance. The point at which this occurs is contingent upon the time between the last study of the material and the final test or recall event (i.e., the retention interval).

Relative to shorter lags, a longer lag where more material is encountered between each presentation of a certain piece of information, enhances memory performance (Wissman et al., 2012). JOLs are often either not sensitive to the memorial benefits of a longer lag, or fail to recognise the disadvantages of shorter lags (Cohen, Yan, Halamish, & Bjork, 2013; Logan et al., 2012; Pyc & Rawson, 2012). This means that as performance increases with longer lag times, there is no concomitant rise in JOLs. Alternatively, people may be liberal in the assignment of JOLs in the short lag, even though there is no commensurate increase in memory. In one study, approximately 70% of participants believed that smaller lags resulted in superior memory for material (Wissman et al., 2012). In addition, other research has revealed poorer performance, yet greater JOLs for a short lag relative to longer lags (Pyc & Rawson, 2012). This suggests people may not be sufficiently aware of how much (or little) they are learning when there are few intervening items compared to several intervening items.

Both criterion learning and lag reflect conditions of learning and are thus considered extrinsic cues for metacognitive judgments. For example, the number of times information is encountered has a direct positive impact on JOLs and learning (Pyc & Rawson, 2012). While Koriat (1997) initially suggested that over time

extrinsic cues are discounted in favour of mnemonic cues, Pyc and Rawson (2012) proposed that metacognitive beliefs regarding retrieval practice can be influential. Specifically, when learning to criterion was employed, JOLs increased over successive trials and this was partly based on criterion level as an extrinsic cue. In contrast, if any beliefs exist regarding lag, they seem to be dismissed, as Koriat proposed and recognised in the literature (Cohen et al., 2013; Logan et al., 2012; Pyc & Rawson, 2012). Moreover, prior exposure to material, as in the case with criterion learning (extrinsic cue), impacts retrieval fluency (mnemonic cue) and therefore JOLs. Karpicke (2009), and Pyc and colleagues (2012) found retrieval fluency impacted JOLs such that faster recall was associated with greater JOLs. For example, Pyc and Rawson found decreased response latencies, a fluency measure, for higher criterion levels and thus increased JOLs at higher levels as well as greater performance. The same result was found for the shorter lag where latency decreased and JOLs increased in comparison to longer lags (Pyc & Rawson, 2012). However, in this instance, longer lags were associated with superior memory performance. Potentially, this may be because information is less accessible in memory, thus leading to greater effort in retrieval (Cull, 2000; Kornell, 2009), which suggests there may have been more scope for potential learning (Vaughn, Hausman, & Kornell, 2017). Furthermore, Logan et al. (2012) posited that for presentations following the initial viewing, longer lag conditions may lead participants to insufficiently adjust their JOLs from the anchor point. It was suggested that this may be due to these items appearing to be less fluently processed; conversely, JOLs for short lag items may unnecessarily increase over time as subsequent presentations are linked to increasingly fluent recall. This may lead to the intuitive, yet incorrect, response of further increasing JOLs which are based on the original anchor that may already be

unnecessarily high, as initial JOLs are often overconfident (Koriat et al., 2002; Rast & Zimprich, 2009).

## Influence of Scale Type on JOLs

More recently, researchers have considered the relationship between the way in which JOLs are measured and metacognitive accuracy with the proposition that inaccuracy is partly an artifact of the scale type (Dunlosky et al., 2005; Mueller, Dunlosky, & Tauber, 2015; Zawadzka & Higham, 2015). JOLs are commonly measured on the 0 (definitely will not be able to recall the information) – 100% (definitely will be able to recall the information) scale described earlier. Alternatively, a binary system may be used, such that when someone predicts they will recall an item they respond 'yes', and respond 'no' to items they predict they will not recall (Zawadzka & Higham, 2015).

The first point of discussion concerns the 0-100% scale. There are two possible interpretations; probability and confidence, with only the former leading to the possibility of true metacognitive inaccuracy (Hanczakowski et al., 2013). Hanczakowski et al. (2013) proposed that 0-100% scale JOLs are not based on the probability of recalling items, as most researchers assume, but instead reflect one's confidence in their recallability. As such, values on this scale lose their objectivity as JOLs are instead based on rank order (Zawadzka & Higham, 2016). Essentially items are assigned JOLs based on evidence for recall, rather than as a prediction of the proportion of items that will be correctly recalled. Thus, there is merely greater evidence for an item receiving 90% than 80% and it is not the case that one assigns 90% based on believing they will recall about 90% of the items assigned this value. Consequently, over- or underconfidence does not reflect true cognitive bias. Hanczakowski and colleagues have claimed binary judgments more accurately

measure metacognitive beliefs as they better index subjective probability.

Regarding the anchoring hypothesis, future 0-100% scale judgments increase in inaccuracy as they are biased towards the anchor value (Tversky & Kahneman, 1974). Overall, if JOLs are based on an initial anchor point, then a probability interpretation means participants are truly over- or underconfident. Here, absolute metacognitive accuracy becomes impaired over multiple trials with 0-100% scales. However, under the confidence interpretation this means inaccuracy is only an artifact of the measurement scale and does not reflect true cognitive distortion. When the 0-100% scale is perceived in terms of confidence, the anchor point itself becomes somewhat arbitrary. This interpretation has some support in the literature in that binary measures diminished metacognitive inaccuracy, thus indicating people accurately monitored their learning over multiple study-test trials (Hanczakowski et al., 2013). As a result, binary judgments are considered less sensitive to the anchor point due to the polarised measurement format and are suggested to be a superior representation of subjective probability (Zawadzka & Higham, 2015).

## Current Study

It is well established that study strategies including practice testing and distributed practice are highly effective (Dunlosky et al., 2013). However, it is poorly understood how people make metacognitive judgements for material they know well, such as when learning to criterion. Therefore, the current research aimed to investigate criterion learning and lag, as well as the impact of the type of JOL scale used, in order to best help people gauge their level of learning and understanding of material for which memory is strong, including whether under- or overconfidence would emerge and under which conditions. Criterion learning and lag have received considerably less attention than has the broader practice testing and distributed

practice paradigms, thus warranting investigation into these valuable study strategies.

Overall, this novel combination of variables provides valuable insight into their influences within a metacognitive framework. This is crucial as it is important to understand whether people have sufficient metacognitive awareness, as the literature generally suggests otherwise, in order to reduce the likelihood of inadequate learning or superfluous overlearning (Rast & Zimprich, 2004). Consequently, attempts to improve this awareness become necessary, unless any inaccuracy can be accounted for by the JOL scale. Moreover, this information is not only useful to those who are learning the material (i.e., students), but for teachers and instructors also, as it has been found that they too often lack insight into robust study techniques (Morehead, Rhodes, & DeLozier, 2016).

Based on the literature reviewed several predictions were put forward. Consistent with the literature (e.g., Pyc & Rawson, 2009), it was expected that superior cued recall performance would emerge as criterion level increased. It was expected that higher JOLs would be assigned across criterion levels, on the basis of participants using criterion as an extrinsic cue. As one must work harder in the long lag (Cull, 2000), relative to the short lag, it was expected that the long lag would lead to enhanced recall capacity, but this may not be reflected by the magnitude of JOLs assigned. Should participants possess insight into the memorial benefits of a longer lag, they would assign higher JOLs here relative to the short lag. However, prior research (Pyc & Rawson, 2012) strongly suggests people do not appreciate the benefits of the longer lag, thus it was expected that JOLs would not display this pattern. Instead, based on previous findings, JOLs in the long lag condition were expected to be lower or similar to those in the short lag condition. As a result, JOLs were expected to be overconfident in the short lag.

However, based on Hanczakowski et al.'s (2013) research, it was expected that the above pattern may differ between continuous and binary JOLs. If binary JOLs tap into probability judgments about the likelihood of recall to a greater extent than continuous JOLs, then participants may be less susceptible to mispredicting the effects of lag on memory. As such binary JOLs may make participants more sensitive to the memorial benefits of the longer lag. If this is the case, it would be expected that the effects of lag on JOLs would more closely correspond to lag effects on memory. Consequently, JOLs would be higher in the long lag compared to the short lag, and JOLs would be less overconfident in the short lag, as compared to when the continuous scale was used.

## Method

### Participants

A total of 46 participants (27 female) took part in the experiment, ranging in age from 18-63 years ($M$ = 27.91 years, $SD$ = 11.46). Participants were recruited from the University of Tasmania, Flinders University, and wider community. It was a requirement that participants not have knowledge of Lithuanian. For their time, participants received $40.00 or were awarded course credit.

A G*Power analysis determined a minimum sample size of 28 was required to detect a moderate effect (Cohen's $f$ = .25; Cohen, 1998), with alpha set at .05 and power at .95. This, in conjunction with satisfying the benchmark recommended by Simmons, Nelson, and Simonsohn (2011), of 20 participants per level of the between-subjects variable indicated the appropriateness of the current sample and that the study had adequate power.

**Materials and Procedure**

**Overview.** The study encompassed three separate phases, as depicted in Figure 1. Briefly, phase 1 involved participants studying Lithuanian-English word pairs. Phase 2 included practice testing of these word pairs until they were successfully recalled a specified number of times, according to the criterion for the given pair, as will be described in the practice test-restudy section of the Method. Phase 3 contained a final cued recall task for all the word pairs. All major elements of these stages were computer-administered using custom programs via Java software for the first two phases and LimeSurvey software for the remaining phase.

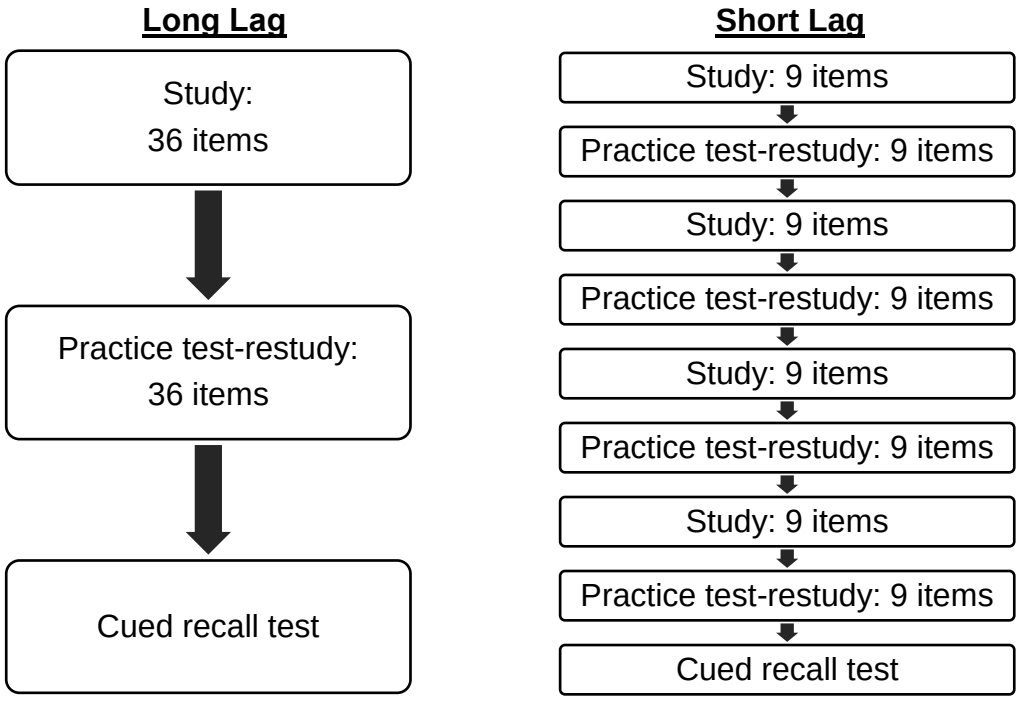


*Figure 1*. Phases of the experimental procedure.

**Study phase.** After obtaining informed consent (see Appendices B and C), participants were informed that they would be required to learn Lithuanian-English word pairs that they would subsequently be asked to remember in a cued recall task. It was stated that during the study some screens automatically progressed while others would require the participant to click a 'continue' button to proceed. All

further instructions were displayed to participants on the computer screen (refer to Appendix D). Before the instructions regarding the presentation of the first list of items were displayed, participants were asked to complete a set of demographic questions. Included here was a question confirming that the participant was unable to speak Lithuanian.

Participants were informed that the experiment would begin with them studying a list of Lithuanian-English word pairs, individually displayed on screen for 6s, followed by several test-restudy trials. A total of 72 Lithuanian-English word pairs (e.g., langas-window; see Appendix E) previously normed for difficulty (Grimaldi, Pyc, & Rawson, 2010) were included. Lithuanian was chosen due to its reasonably uncommon status as a Baltic language which diverges from more common romance languages (e.g., French; Grimaldi et al., 2010). Furthermore, it uses the English alphabet, thus removing any translation complications.

Conforming to the procedure followed by Pyc and Rawson (2012), the word pairs were randomly allocated to one of two lists and equivalent item difficulty was ensured. The lists were counterbalanced across lag conditions, the order of which was also counterbalanced across participants. This meant that each lag was associated with each list an approximately equal number of times and that half the participants received the short lag followed by the long lag, with the other half receiving the long lag followed by the short lag.

As shown in Figure 1, for the long lag condition, participants were presented with all 36 items from a single list in the initial study phase, before proceeding to the next phase. For the short lag condition, the 36 items of a single list were randomly assigned to four lists. Therefore, as shown in Figure 1, phases 1 and 2 where interleaved. Hence, each list comprised nine word pairs with equivalent item

difficulty in each. Participants received the initial study phase for the first list of nine word pairs prior to the next phase (to practice that list only), before again returning to the study phase for the second list. The order of presentation of the four lists in the short lag was counterbalanced across participants. The above procedure ensured an equal number of word pairs were studied in each lag condition, while manipulating the number of intervening items between presentations of each pair.

**Practice test-restudy.** After a list of word pairs had been presented in the study phase, participants were informed that they would be tested on these. Participants were asked to correctly identify the English target of the studied word pair when the Lithuanian cue appeared on the screen. Responses were scored correct if the first five letters matched the target word; this ensured responses were clearly distinguishable as the target word before being scored correct (see Frank & Kuhlmann, 2017). Participants were given 8s to respond correctly; if they did not respond or responded incorrectly then the item was placed at the end of the list, and the correct answer was presented on screen for a 4s restudy period before proceeding to the next item. When the English target was correctly recalled, there was no restudy period and participants were presented with the next item immediately after the 8s period had elapsed.

The word pairs were continually presented until they were correctly recalled to their specified criterion of either one, three, or nine. For each list, an equal number of items (for each difficulty level) were randomly allocated to each of the criterion levels, with assignment randomised anew for each participant. In line with Pyc and Rawson's (2012) research, participants were unaware of the precise criterion of each item and were told the test-restudy periods would continue until an 'acceptable level of performance' had been reached.

For the long lag condition, word pairs were initially separated by 35 intervening items (the other items in that list). In the short lag, there were initially eight intervening items (the other items in that list). In both conditions once an item was learned to criterion it meant the number of intervening items for the remaining word pairs decreased. This is known as a contracting schedule compared to an equal schedule that has a constant number of intervening items. While this led to both lags essentially becoming shorter over time, there is evidence that this is not problematic since it is the absolute spacing of items that is important (Karpicke & Bauernschmidt, 2011). That is, the sum of all intervening items between each presentation of a given word pair, rather than relative spacing (i.e., the number of intervening items between any two presentations of a given word pair). As such the absolute spacing would inevitably be greater in the long lag condition as it consisted of more items.

For both lag conditions, immediately after an item was learned to its predetermined criterion level, participants were asked to provide a JOL for that word pair. Half the sample were randomly allocated to each scale type. In the 0-100% condition, participants were asked to indicate how likely they would be to recall the English word when presented with the Lithuanian word alone, on a test about 7 days later. The response options available to participants were in 10% increments (i.e., 0%, 10%, …, 100%). In the binary condition, participants were asked to indicate whether they were likely to recall the English word when presented with the Lithuanian word alone, on a test about 7 days later. Participants answered by responding either ‘yes’ (*I think I will be able to recall this item*), or ‘no’ (*I do not think I will be able to recall this item*). Participants had unlimited time to respond before moving to the next item.

Once word pairs in a given list were recalled to criterion, participants started the study phase again for the next list. This being either the first list of the short lag, if participants completed the long lag first, or the second list of the short lag if they started with the short lag condition. In total, participants repeated the study and practice test-restudy phases five times.

**Final cued recall task.** Following the completion of the first two phases, participants returned 5 to 10 days later ($M$ = 6.7 days, $SD$ = 1.2) to complete the final cued recall task. The self-paced test included all 72 word pairs which were presented in random order to participants, irrespective of the lists in which they first received the items. All items were presented individually. Responses were considered correct if they were an exact match to the target or plurals of the target, this essentially reflected the criteria in the practice test-restudy phase, with the exception that obviously misspelled words were also considered correct. As with the practice testing phase, the Lithuanian cue was presented alone, and participants were required to respond with the English target. If participants did not know the answer, or did not wish to attempt a response, they could type an 'X' in the response box to continue to the next word pair. Participants had one attempt at each item and were not informed whether they had correctly recalled the English word or not.

## Design

The experimental design of the current study was a 2 × 2 × 3 mixed factorial design. Within-subjects factors were Criterion (1, 3, 9) and Lag (short, long). The between-subjects factor, to which participants were randomly allocated, was Scale (0-100%, binary). The dependent measures were mean JOL ratings, recall accuracy, absolute accuracy, and relative accuracy.

## Results

For pairwise comparisons, reported effect sizes are Cohen's *d*, with the following criteria, 0.20 as a small effect, 0.50 as a moderate effect, and 0.80 as a large effect (Cohen, 1988). Also presented with these are 95% confidence intervals (CI; Cumming, 2012). All other effect sizes reported are Cohen's *f*, with the following criteria, .10 as a small effect, .25 as a moderate effect, and .40 as a large effect (Cohen, 1988). Bonferroni adjustments were applied to comparisons where necessary, however of note, this did not alter the interpretation of any results.

Prior to analysis, data were examined to determine whether they met the assumptions required. Relative accuracy data were positively skewed, therefore square root transformations and adjustment of extreme scores to one unit above the next closest score were applied. These solutions either did not improve the distribution or did not produce different results to analyses conducted with untransformed data, thus for simplicity in interpretation, analyses using untransformed data are reported. There were no problematic outcomes for the other variables. Additionally, analyses affected by sphericity being violations, as identified by significant Mauchly's statistics, are reported with a Huynh-Feldt[1] correction. Where sphericity was not violated, or the variable comprised only two levels, no corrections were applied.

Preliminary analyses indicated no significant differences in the 0-100% and binary scale groups regarding mean age, sex distribution, or retention interval (all *t*s and $\chi^2 < 1$).

[1] Huynh-Feldt corrections were chosen over Greenhouse-Geisser corrections as estimates of sphericity were > .75, thus favouring the less conservative option.

**Recall Accuracy**

One objective of the present research was to investigate the factors affecting recall accuracy. To accomplish this a 2 × 2 × 3 (Lag [long, short] × Scale [0-100%, binary] × Criterion [1, 3, 9]) mixed analysis of variance (ANOVA) was conducted.[2] The main effect of criterion was significant, $F(2, 88) = 67.31$, $p < .001$, $f = .496$. Consistent with the literature, this showed that learning items to a higher criterion was beneficial for memory, as shown in Figure 2.

The memorial benefits of a longer lag have previously been established. This finding was replicated in the current study as indicated by the significant main effect of lag (see Figure 2 for means), $F(1, 44) = 84.82$, $p < .001$, $f = .566$.

The above main effects were qualified by a significant interaction (see Figure 2). This indicated the benefits of learning to a higher criterion were greater in the long lag compared to short lag condition, $F(1.74, 76.51) = 21.06$, $p < .001$, $f = .219$.[3] To demonstrate this, Bonferroni adjusted paired samples *t*-tests were conducted between adjacent criterion levels for each lag. For the long lag, recall accuracy was greater at criterion 3 than criterion 1, $t(45) = 6.53$, 95% $CI_{difference}$ [11.27, 21.33], $p < .001$, $d = 0.82$, 95% CI [0.48, 1.15]. Accuracy was also greater at criterion 9 compared to criterion 3, $t(45) = 5.10$, 95% $CI_{difference}$ [8.77, 20.22], $p < .001$, $d = 0.59$, 95% CI [0.28, 0.91]. The short lag followed a similar pattern, though performance was enhanced to a smaller degree between successive levels. Thus, recall accuracy was higher at criterion 3 than criterion 1, $t(45) = 2.58$, 95% $CI_{difference}$ [1.12, 9.03], $p = .013$, $d = 0.33$, 95% CI [0.03, 0.62]. Accuracy was also higher at criterion 9

[2] Results from this analysis not reported in the following sections are displayed in Appendix F, Table F1.

[3] Mauchly's test for criterion × lag interaction, $\chi^2(2) = 7.00$, $p = .030$, $\tilde{\varepsilon} = .923$.

compared to criterion 3, $t(45) = 2.61$, 95% $CI_{difference}$ [1.20, 9.31], $p = .012$, $d = 0.28$, 95% CI [-0.01, 0.58].

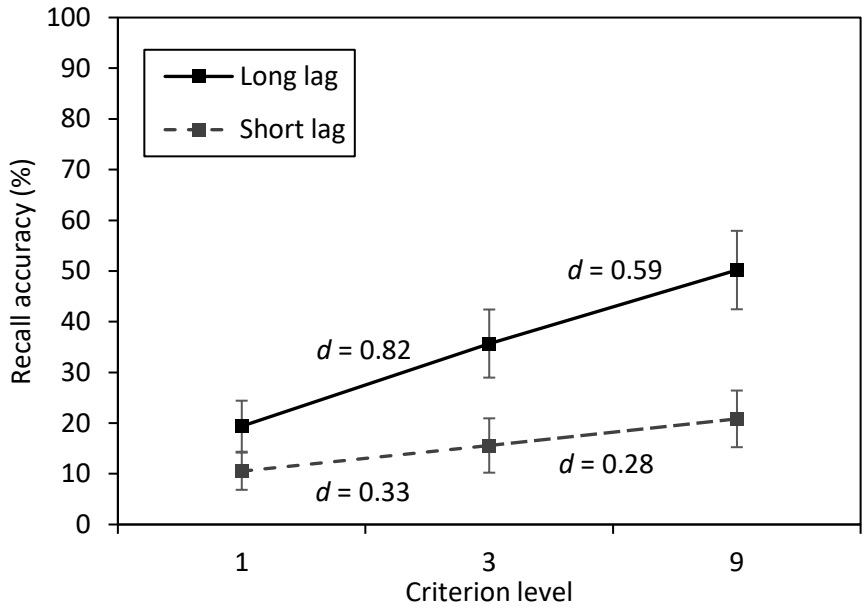


*Figure 2.* Recall accuracy for the long lag and short lag conditions at each criterion. Error bars represent 95% confidence intervals. Cohen's *d* values refer to the degree of increased recall accuracy across successive criteria.

**Metacognitive Judgments (JOLs)**

To analyse whether the predicted effect regarding mean JOLs occurred, a 2 × 2 × 3 (Lag [long, short] × Scale [0-100%, binary] × Criterion [1, 3, 9]) mixed ANOVA was conducted. As expected, the overall effect of criterion on mean JOLs was significant, $F(1.52, 66.72) = 70.58$, $p < .001$, $f = .427$.[4] Pairwise comparisons revealed significant differences between each criterion level (all *p*s < .001), as such JOLs were significantly greater at criterion 9 ($M = 60.00$, $SD = 29.61$, 95% CI [51.20, 68.78]) than criterion 3 ($M = 45.00$, $SD = 28.68$, 95% CI [36.48, 53.52]),

[4] Mauchly's test for criterion main effect, $\chi^2(2) = 16.52$, $p < .001$, $\tilde{\varepsilon} = .797$.

which in turn were significantly greater than criterion 1 ($M$ = 31.20, $SD$ = 24.32, 95% CI [23.98, 38.43]). Thus, JOL ratings increased as a function of increasing criterion.

Although the difference was not statistically significant, numerically higher mean JOLs were reported in the short lag ($M$ = 48.32, $SD$ = 27.99, 95% CI [40.00, 56.63]) compared to the long lag ($M$ = 42.48, $SD$ = 28.13, 95% CI [34.13, 50.84]), $F(1, 44) = 3.31$, $p = .076$, $f = .126$. There were no other significant findings in this analysis, and given they were not pertinent to the research aims they are not included here.[5]

**Absolute Accuracy of JOLs**

To assess the correspondence between recall accuracy and JOLs, a 2 × 2 × 2 × 3 (Lag [long, short] × Scale [0-100%, binary] × Measure [JOLs, accuracy] × Criterion [1, 3, 9]) mixed ANOVA was conducted. In order to perform this analysis, we followed a procedure used by Koriat (1997; see also Hanczakowski et al., 2013; Logan et al., 2012). This involved including JOLs and recall accuracy as a within-subjects factor, labelled 'measure', in the ANOVA. This procedure permits continuous and binary JOLs to be compared on equal grounds. As such, binary JOLs were first converted to produce mean percentage JOLs, whereby 'no' JOLs became 0%, while 'yes' JOLs became 100%, a method substantiated in the literature (Zawadzka & Higham, 2015).

The ANOVA revealed a significant main effect of measure, $F(1, 44) = 24.71$, $p < .001$, $f = .481$, and significant interactions between criterion and measure, $F(1.67, 73.52) = 5.00$, $p = .013$, $f = .076$,[6] as well as lag and measure, $F(1, 44) = 53.92$, $p <$

[5] In the interest of comprehensive reporting these results are displayed in Appendix F, Table F2.

[6] Mauchly's test for criterion × measure interaction, $\chi^2(2) = 13.10$, $p = .001$, $\tilde{\varepsilon} = .835$.

.001, $f$ = .279. These findings were qualified by the interaction reported below. Results from this analysis that were not statistically significant can be found in Appendix F (Table F3) as they are not integral to the research objectives and are superseded by the following key interaction .[7]

As depicted in Figures 3 and 4, the interaction between criterion, lag, and measure was significant, $F(2, 88) = 6.91$, $p = .002$, $f = .122$. To further examine this interaction, separate 2 × 3 (Measure [JOLs, binary] × Criterion [1, 3, 9]) repeated measures ANOVAs were conducted for each lag condition.

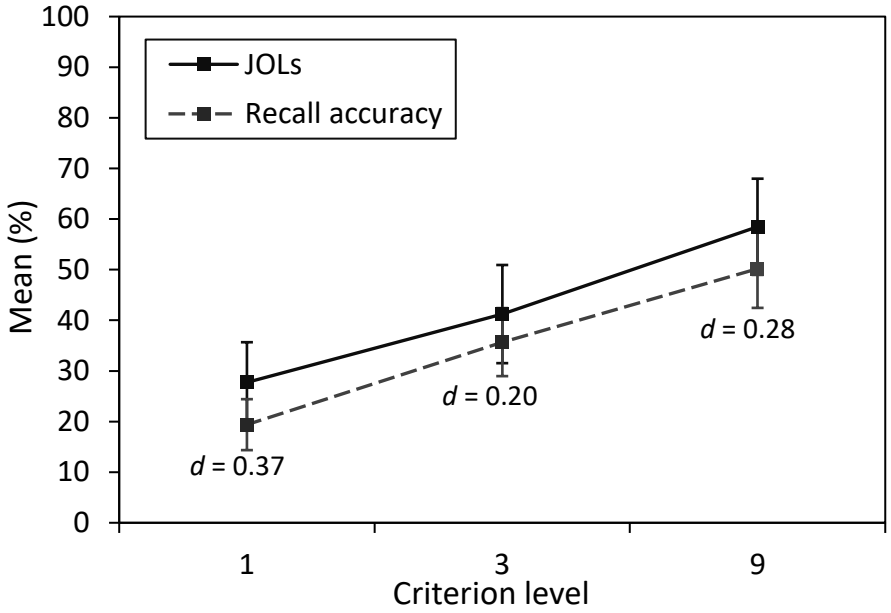


*Figure 3.* Mean JOLs and recall accuracy in the long lag condition at each criterion level. Error bars represent 95% confidence intervals. Cohen's *d* values reflect magnitude of overconfidence at each criterion.

[7] For analyses containing the measure variable, any results not involving this variable are not reported since it is futile to collapse across measure.

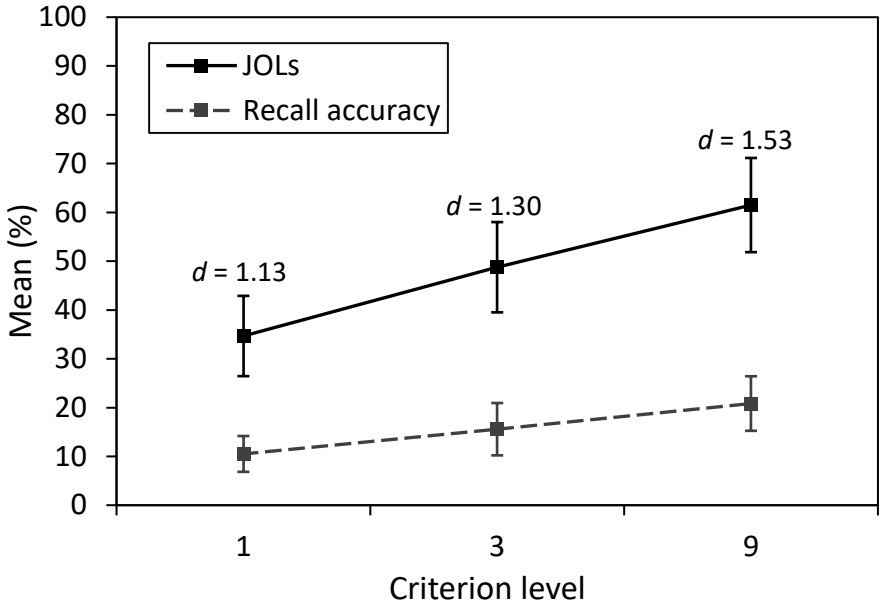


*Figure 4.* Mean JOLs and recall accuracy in the short lag condition at each criterion level. Error bars represent 95% confidence intervals. Cohen's *d* values reflect magnitude of overconfidence at each criterion.

**Long lag.** The main effect of measure was not significant, $F(1, 45) = 2.80$, $p = .101$, $f = .155$. The interaction between criterion and measure also was not significant, $F(2, 90) = 0.42$, $p = .661$, $f = .025$, indicating that JOLs and recall accuracy both increased at approximately the same rate over criterion levels. To further demonstrate this, paired samples *t*-tests were conducted at each of the criterion levels (see Figure 3 for descriptive statistics).[8] There were no significant differences between JOLs and recall accuracy at criterion 1, $t(45) = 1.90$, $p = .064$, $d = 0.37$, 95% CI [0.07, 0.67]; criterion 3, $t(45) = 1.09$, $p = .284$, $d = 0.20$, 95% CI [-0.10, 0.49]; or criterion 9, $t(45) = 1.64$, $p = .108$, $d = 0.28$, 95% CI [-0.01, 0.58]. The effect sizes associated with these comparisons were small, suggesting a relatively minor degree of overconfidence compared to the short lag.

[8] For full transparency, specific descriptive statistics are displayed in Appendix F, Table F4, including statistics for the following short lag analysis.

**Short lag.** There was a significant main effect of measure such that JOLs were greater than recall accuracy, $F(1, 45) = 58.40$, $p < .001$, $f = .763$. However, this was qualified by a significant interaction between criterion and measure, $F(1.67, 74.91) = 11.43$, $p < .001$, $f = .144$,[9] indicating that unlike the long lag, JOLs and recall accuracy increased to a dissimilar degree across criterion levels. To explore this finding paired samples *t*-tests were conducted at the individual levels of criterion (see Figure 4 for descriptive statistics). Overconfidence was observed at criterion 1, as such JOLs were significantly higher than recall accuracy, $t(45) = 5.52$, 95% $CI_{difference}$ [15.35, 32.99], $p < .001$, $d = 1.13$, 95% CI [0.75, 1.49]. The same was result was found at criterion 3, though to a greater degree, $t(45) = 7.29$, 95% $CI_{difference}$ [24.02, 42.35], $p < .001$, $d = 1.30$, 95% CI [0.91, 1.70]. Again this was the case for criterion 9, but to a greater extent still, $t(45) = 7.84$, 95% $CI_{difference}$ [30.22, 51.12], $p < .001$, $d = 1.53$, 95% CI [1.10, 1.96]. Thus, in the short lag condition, JOLs became more overconfident as criterion increased.

**Relative Accuracy of JOLs**

The Adjusted Normalised Discrimination Index (ANDI) is a measure of relative metacognitive accuracy and as such it assesses the extent to which JOLs differentiate items that will be correctly and incorrectly retrieved (Yaniv et al., 1991). Values on the ANDI statistic range from 0 (poorest resolution) to 1 (perfect resolution).

In order to assess discrimination ability, a 2 × 2 × 3 (Lag [long, short] × Scale [0-100%, binary] × Criterion [1, 3, 9]) mixed ANOVA was conducted. The significant main effect of lag indicated that relative accuracy was enhanced in the long lag ($M = .12$, $SD = .12$, 95% CI [.08, .16]) compared to the short lag ($M = .03$,

[9] Mauchly's test for criterion × measurement interaction, $\chi^2(2) = 12.05$, $p = .002$, $\tilde{\varepsilon} = .832$.

*SD* = .07, 95% CI [.02, .04]), $F(1, 44) = 14.47$, $p < .001$, $f = .413$. While the following difference was not significant, the main effect of scale suggests there may be a benefit of continuous JOL scales (*M* = .12, *SD* = .12, 95% CI [.06, .17]) compared to binary JOL scales (*M* = .08, *SD* = .02, 95% CI [.03, .12], $F(1, 44) = 3.82$, $p = .057$, $f = .276$. There were no other significant effects found in this analysis, thus refer to Appendix F (Table F5) for these results as they are not crucial to the experimental aims.

## Discussion

The main aim of the current experiment was to investigate the effects of study strategies and JOL scale-type on metacognitive accuracy. A secondary objective was examining the effects of these strategies on memory performance. Several findings were consistent with the hypotheses and literature, while others differed somewhat.

As predicted, superior recall accuracy was found as criterion increased and in the long lag. Specifically, the findings indicate that practice testing is more effective with an increased number of successful retrievals. This is congruent with the criterion learning (Karpicke, 2009; Pyc & Rawson, 2012; Pyc et al., 2014; Vaughn & Rawson, 2011) and broader practice testing (Cull, 2000; Dunlosky et al., 2013) literature. The present findings also indicate that in repeated practice testing scenarios, memory performance benefits from a long lag between the re-testing of items; a finding consistent with research into the lag effect (Cull, 2000; Logan et al., 2012; Pavlik & Anderson, 2005; Wissman et al., 2012) and broader distributed practice paradigm (Kornell & Bjork, 2007).

Of greatest interest in the present research were the effects on metacognition. As in the literature (Pyc & Rawson, 2012), it was hypothesised that higher JOLs would be assigned across increasing criterion levels. The results provide support for

this, which suggests people could appreciate that learning material to a higher criterion would enhance their memory. The current pattern of results across criterion levels are similar to Pyc and Rawson's (2012) findings; thus, potentially supporting their observation of criterion as an extrinsic (belief-based) cue, and the associated retrieval fluency of higher criterion, as influential factors.

As reported in the literature (Cohen et al., 2013), it was expected that the disadvantages of the short lag would be overlooked, thus yielding higher JOLs relative to accuracy. Consequently, this meant short lag JOLs may be similar to or higher than JOLs in the long lag. The current findings indicate similar JOLs across the conditions, with a numerical trend in the posited direction. This adds to the literature suggesting that people do not appreciate that the longer lag benefits memory relative to the short lag; indeed, the JOL patterns indicate that, if anything, the shorter lag leads to higher predictions of recall.

Additionally, poorer metacognitive accuracy was expected in the short lag as demonstrated by Pyc and Rawson's (2012) research, but to a greater degree for those assigning JOLs on the 0-100% scale, as advocated by Hanczakowski and colleagues (2013). The first part of this prediction was supported; the long lag resulted in superior absolute accuracy, however accuracy in both lags was consistent across scales. Specifically, both scale conditions performed poorer in the short lag, both metacognitively and regarding memory performance. This was evidenced by substantial overconfidence for the short lag that was less apparent in the long lag. Discovering the presence of overconfidence (or underconfidence) was an area of interest in conducting this study. Thus, similar to other research (Logan et al., 2012), overconfidence was apparent in the short lag where there were few intervening items, the magnitude of which increased across criterion levels. This increase in magnitude

may stem from participants' failure to acknowledge the diminishing returns of higher criterion levels on future memory performance (Pyc & Rawson, 2012). It is reasonable to expect binary JOLs to assist in overcoming this lack of insight, but there were no meaningful differences between the scales.

The above findings suggest that participants did not seem to understand the effects of lag on performance in that the JOLs assigned in the short lag condition were, if anything, slightly higher than those assigned in the long lag condition, yet the long lag enhanced metacognition in several ways. First, JOLs suggested that people did not believe the long lag would benefit their memory as much as the short lag, however, their JOLs were ultimately much more similar to actual recall than were short lag JOLs. Moreover, the long lag assisted in predicting the likelihood of remembering across criterion levels as the correspondence between recall accuracy and JOLs was maintained.

These current findings do not readily support the confidence interpretation of 0-100% JOLs in either lag condition. Unless there are clear benefits to the use of the binary scale, then the artifact view of the 0-100% scale cannot be supported (Mueller et al., 2015). Moreover, other researchers found binary JOLs did not completely correct metacognitive inaccuracy when compared to continuous scales, thus leading the authors to confirm the influence of other factors such as the deficient application of knowledge (Mueller et al., 2015). This leads to the question of why such differences have been found. One potential reason for the inconsistencies between the present research and that by Hanczakowski et al. (2013) and Zawadzka and Higham (2015) may be that the present study imposed criterion learning, while practice trials where employed in the previous research. A fixed number of trials results in learning status differences (i.e., not all words are correctly recalled an equal

number of times) which can act as a cue for JOLs. In contrast, such differences in learning status cannot be used as a cue for assigning JOLs in criterion learning since all items have necessarily been correctly recalled the specified number of times (Pyc & Rawson, 2012). Essentially, the types of cues used are not homogeneous across these learning techniques. Furthermore, a greater number of correct retrievals enhances recall accuracy, thus leading to potentially differing effects on metacognitive accuracy between the criterion and purely practice trials (Vaughn & Rawson, 2011).

Additionally, the previous scale research utilised three practice cycles at most, with JOLs recorded after each. In contrast, in the current study, JOLs were reported only once the item had reached criterion, and again this may have influenced the cues used by participants. Support for this idea comes from findings in both the current and past research that some overconfidence was shown on criterion or cycle 1 for binary judgments. Beyond this point, it is reasonable to propose that in the present research, over the course of several retrieval attempts, mnemonic or experiential cues increased in their influence on JOLs (Koriat, 1997). Moreover, Serra and Dunlosky (2005) found no such effects of retrieval fluency using a design with two practice cycles. While Zawadzka and Higham (2015) speculate about the impact of fluency cues, especially in relation to several more cycles, they did not examine this. Conversely, Karpicke (2009) and Pyc and Rawson (2012) found retrieval fluency was influential in JOL assignment over increasing criterion levels. Furthermore, differences were found in JOL assignment and accuracy depending on the timing of the judgments. For example, aggregate judgments made after more than three trials were, in one study, much lower in magnitude than JOLs given following the first correct recall (Karpicke, 2009).

Now turning to relative metacognitive accuracy, the current study offers unique insight in that it expands upon the extant metacognitive literature by being the first to investigate the effect of lag. Similar to the predictions for absolute accuracy, poorer relative accuracy was expected in the short lag than the long lag, and potentially to a larger degree for the 0-100% scale than the binary scale. Greater accuracy for binary JOLs was expected, such that increased differentiation between the short and long lag may be found as the dichotomous response format may heighten participants' sensitivity to the lack of benefit from shorter lags. Contrary to the predicted scale effect, the binary condition did not outperform the 0-100% condition and, if anything, the findings indicate the opposite occurred. The results did however provide support for superior accuracy in the long lag in comparison to the short lag.

One explanation for the converse findings regarding the effect of scale is that despite all items being learned to criterion, memory for these will not be identical. Memory will be strong for some items, yet weaker for others (Koriat, 1997), with the stronger items more likely to be recalled on a subsequent test (Kornell, Bjork, & Garcia, 2011). When making JOLs, people might have some capacity to distinguish items with a stronger memory from those with a weaker memory and might give, on average, different levels of JOLs to each. For example, one may assign 90% JOLs to items they perceive as having a stronger memory for and 60% for weaker items. Thus, if more of the stronger items than the weaker items are recalled on the test, this indicates good discrimination: the individual was able to use JOLs to distinguish between items they would remember and items they would not.

In contrast, for binary JOLs, one may still have some insight into the strength of memory for different items, but when asked to make JOLs on a binary scale this

insight might not translate into differences in JOLs. For example, stronger items and weaker items may all be assigned mostly ‘yes’ JOLs. The only alternative response is ‘no’, therefore offering a limited margin of error; thus, binary scales may constrain JOL assignment (Serra & Ariel, 2014). If this is the case, then discrimination would be poorer than when continuous JOLs are used. These findings suggest dichotomous judgments do not benefit people in deciding what they do and do not know when the criterion learning paradigm is applied in combination with lag.

In the present research, anchoring may account for the increased JOLs across increasing criterion levels. Participants started at a point of 30-40% (i.e., early criterion 1 items in the long lag and from the first list of the short lag), in line with the literature (Rast & Zimprich, 2009). Participants may have used this as a beginning mark and adjusted JOLs upward on the basis of believing higher criterion levels would be better for memory or due to retrieval fluency effects (Logan et al., 2012). Alternatively, the *memory for past test heuristic* (Finn & Metcalfe, 2008) provides a plausible explanation of the outcomes. Here, memory for previous trials may be relied upon in assigning JOLs, with success on previous trials associated with higher JOLs (Ariel & Dunlosky, 2011). This could account for the apparent overconfidence whereby participants thought about the last recall attempt, which was necessarily correct in the present research.

There are numerous practical implications based on the current findings regarding the way students study, particularly self-regulated study. The outcomes provide further support for criterion learning in enhancing memory, and as such indicates we should promote the use of this strategy among students. Particular focus should be on encouraging the application of higher rather than lower criterion levels (Pyc & Rawson, 2009).

Additionally, longer lags will likely benefit students, and given participants in the current and previous studies (Cohen et al., 2013; Pyc et al., 2014) failed to recognise the disadvantages of the shorter lag, it seems most valuable to raise awareness of this. Moreover, given how commonly 'cramming', or the intense study of material in a short period prior to a test, is engaged in (Blasiman, 2017; Gerbier & Toppino, 2015; Kornell, 2009; Son & Kornell, 2009), it seems prudent to promote spaced learning. One suggestion put forward (Delaney, Verkoeijen, & Spirgel, 2010) is to prompt teachers or instructors to foster spaced learning, especially in a way that encourages self-regulated use.

Moreover, this study provides preliminary evidence that a longer lag not only enhances memory but also provides a condition under which people more accurately monitor their learning (i.e., what they are likely to remember on a future task). This has clear applied value in that it provides a reasonably accurate indication of one's understanding of the information, and thus, potentially how much additional time they need to allocate to that material. This also highlights the need to help people better monitor their learning should they be required to study or learn at a shorter lag, or ideally, assist them in understanding that this approach is generally unhelpful, and guide them toward more beneficial techniques.

**Limitations and Future Directions**

While providing new insight, the current study had potential limitations. First, for the 0-100% scale, options were in 10% increments meaning full variation was not captured. While a broader range of responses were available in this condition relative to binary JOLs, even more fine-grained distinctions were not possible. Mickes, Wixted, and Wais (2007) found greater variability in JOLs when more options were available. However, even when participants were given 99 options (i.e., 1-99), JOLs

were frequently assigned at 5% intervals, therefore, essentially creating a 20-point scale. While it is not clear whether this finding extrapolates to the present research, it appears more options may have provided greater opportunity to differentiate items. This has ramifications for relative accuracy, in particular, as additional options may allow greater variability in discrimination.

Further in relation to scale, while there are precedents in the literature (e.g., Hanczakowski et al., 2013), the accuracy of converting JOLs from a binary yes/no format to that of 0/100% may be questionable. In an attempt to investigate this, we found re-running the analyses with the continuous scale coded as binary (i.e., JOLs $\leq$50 were assigned 0 and JOLs over this assigned 100) did not alter the outcomes. This provides some statistical evidence for the appropriateness of the binary conversion, however, conceptually, the question remains open.

Another possible limitation relates to whether the outcomes generalise to other memory tasks, such as recognition tests, tasks with larger individual components (e.g., concept definitions), and importantly, in classroom settings (Delaney et al., 2010; Son & Simon, 2012). It has been suggested that the ecological validity of commonly imposed experimental tasks, such as learning word pairs in a brief study period may not be tapping the same memory processes as the tasks students complete over days or weeks in real contexts (Rohrer, 2015). However, the current findings provide a promising foundation. Also, the effects of distributed practice, in general, are somewhat robust across materials (Dunlosky et al., 2013), and practice testing in a broad sense has generalised to educationally relevant stimuli (McDaniel, Roediger, & McDermott, 2007). However, it must be determined how the effects of criterion and particularly lag influence metacognitive accuracy in educational settings.

One objective of the current research was to improve sensitivity to the effects of lag on memory performance. In the current study the use of binary JOL assignment was not beneficial in this regard. Another avenue for future research is examining the effects of lag when using delayed JOLs. Delayed judgments are reported as enhancing one's ability to accurately monitor learning, as reflected by increased JOL accuracy (Dunlosky et al., 2005; Nelson, Narens, & Dunlosky, 2004; Rhodes, 2016). This improvement, known as the *delayed JOL effect*, has been shown to enhance JOL accuracy for criterion learning (Pyc et al., 2014), as well as for distributed compared to massed learning (Dunlosky & Nelson, 1994). Crucially, it remains an open question of how the accuracy of delayed JOLs differs across varying lags.

There are several cues that contribute to JOL assignment and the resultant metacognitive accuracy. Future research looking at these may be highly beneficial in understanding the relationship between study strategies and metacognitive accuracy. A particularly valuable line of inquiry follows the effects of these cues on the impact of lag and resolution. Thus, investigation of these cues would provide insight into how we could attempt to improve metacognitive accuracy.

Therefore, it is worthwhile to consider the influences of belief-based cues in comparison to experience-based or mnemonic cues (Frank & Kuhlmann, 2017), such as fluency or trials to criterion (i.e., how many attempts one requires to reach the specified criterion). There is evidence of trials to criterion influencing both metacognitive and recall accuracy (Pyc & Rawson, 2012). Thus, it would be pertinent to assess which factors influence the interpretation of trials to criterion, such as the individual's beliefs and theory of intelligence. For example, entity theorists (those who believe intelligence is fixed; Dweck, 1999), tend to believe that

extended effort reflects poorer ability and lower likelihood of future remembering. In contrast, incremental theorists (those believing intelligence is malleable; Dweck, 1999), attribute extended effort to high engagement with a task, and are less likely to be discouraged by the number of trials taken as they focus on the potential gains in knowledge through sustained effort. Thus, for items that take considerable time to reach criterion, entity theorists may assign low JOLs and incremental theorists may assign higher JOLs.

Furthermore, it is likely other belief- and experience-based cues are influential, though it may be to varying degrees depending on the task or context. For example, Mueller and Dunlosky (2017) found beliefs about ease of processing given information had a substantial impact on JOLs, while Frank and Kuhlmann (2017) reported quite the opposite in that beliefs regarding the effects of a stimulus (amplitude of a tone) on learning, were insufficient for assigning JOLs and that experience-based cues (e.g., those arising during task completion) were a major influence.

Ideally, the current findings would generalise to real-world contexts, however, as aforementioned, we cannot assume this will be the case. Thus, further research is required to determine how the current findings translate and can best be applied in classroom settings, or at the very least, to other memory or learning tasks. Additionally, it is unlikely that all material needs to be studied to the same criterion in order to be equally well remembered. As some authors (Rawson & Dunlosky, 2011) suggest, people should not pick an arbitrary number of times to study material. To make the most of criterion learning (i.e., studying information to an optimal criterion level), it is likely that people need accurate metacognitive monitoring to

decide which material to assign to which criterion. Again, this underscores the importance of improving metacognitive monitoring.

In conclusion, the current study replicated the effects of a higher criterion and longer lag enhancing memory performance. Again, alluding to the utility of these strategies in educational contexts, especially if further research is conducted in more ecologically valid settings (Rohrer, 2015). Moreover, not only did memory benefit from the long lag, metacognitive accuracy also improved, such that overconfidence was substantially reduced compared to the short lag. The novel finding of increased relative accuracy for the long lag is informative as students often base their study decisions on JOLs (Metcalfe & Finn, 2008), thus highlighting the importance of such judgments for successful learning. The findings also provide support for the claim that people have some accurate understanding of the influence of criterion as higher JOLs were assigned across the criterion levels, with varying degrees of accuracy. However, in contrast to prior research (Hanczakowski et al., 2013; Zawadzka & Higham, 2015), those using the binary scale did not display superior metacognitive accuracy in comparison to those using the continuous scale, thus, one cannot conclude the 0-100% scale does not reflect subjective probability. It appears that this may be limited to the context of certain phenomena, such as the UWP effect. Therefore, future research should continue in its quest to enhance metacognitive accuracy.

# References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*, 171-184. doi: 10.3758/s13421-010-0002-y

Blasiman, R. N. (2017). Distributed concept reviews improve exam performance. *Teaching of Psychology, 44*, 46-50. doi: 10.1177/0098628316677646

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013). Do students think that difficult or valuable materials should be restudied sooner rather than later? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1682-1696. doi: 10.1037/a0032425

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50-71. doi: 10.1037/0882-7974.12.1.50

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215-235. doi: 10.1002/(sici)1099-0720(200005/06)14:3<215::aid-acp640>3.0.co;2-1

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53*, 63-147. doi: 10.1016/S0079-7421(10)53003-2

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*, 545-565. doi: 10.1006/jmla.1994.1026

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4-58. doi: 10.1177/1529100612453266

Dunlosky, J., Serra, M., Matvey, G., & Rawson, K. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology*, *132*, 335-346. doi: 10.3200/GENP.132.4.335-346

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development.* Philadelphia, PA: Psychology Press.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*, 19-34. doi: 10.1016/j.jml.2007.03.006

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*, 567-586. doi: 10.1007/s10648-015-9313-7

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34,* 906-911. doi: 10.1037/0003-066x.34.10.906

Frank, D. J., & Kuhlmann, B. G. (2017). More than just beliefs: Experience and beliefs jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 680-693. doi: 10.1037/xlm0000332

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience and Education, 4*, 49-59. doi: 10.1016/j.tine.2015.01.001

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian-English paired associates. *Behavior Research Methods, 42*, 634-42. doi: 10.3758/BRM.42.3.634

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect, *Journal of Memory and Language, 69*, 429-444. doi: 10.1016/j.jml.2013.05.003

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469-486. doi: 10.1037/a0017341

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1250-1257. Doi: 10.1037/a0023436

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162. doi: 10.1016/j.jml.2006.09.004

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217-273. doi: 10.1016/0001-6918(91)90036-Y

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349-370. doi: 10.1037/0096-3445.126.4.349

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52,* 478-492. doi: 10.1016/j.jml.2005.01.001

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147-162. doi: 10.1037/0096-3445.131.2.147

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23,* 1297-1317. doi: 10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated learning. *Psychonomic Bulletin & Review, 14,* 219-224. doi: 10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits – and costs – of dropping flashcards. *Memory, 16,* 125-136. doi: 10.1080/09658210701763899

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85-97. doi: 10.1016/j.jml.2011.04.002

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval

processes in the lag effect: A multinomial processing tree analysis. *Memory, 20*, 37-47. doi: 10.1080/09658211.2011.631550

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning, 7,* 175-195. doi: 10.1007/s11409-012-9090-3

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200-206. doi: 10.3758/bf03194052

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174-179. doi: 10.3758/PBR.15.1.174

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140*, 239-257. doi: 10.1037/a0023007

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*, 858-865. doi: 10.3758/bf03194112

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory, 24*, 257-71. doi: 10.1080/09658211.2014.1001992

Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language, 93*, 245-258. doi: 10.1016/j.jml.2016.10.008

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition, 43*, 180-192. doi: 10.3758/s13421-014-0474-2

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods, 9*, 53-69. doi: 10.1037/1082-989x.9.1.53

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29*, 559-586. doi: 10.1207/s15516709cog0000_14

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437-447. doi: 10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2012). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition, 40*, 976-988. doi: 10.3758/s13421-012-0200-x

Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: When and why are judgments accurate? *Memory & Cognition, 42,* 886-897. doi: 10.3758/s13421-014-0403-4

Rast, P., & Zimprich, D. (2009). Age differences in the underconfidence-with-practice effect. *Experimental Aging Research, 35*, 400-431. doi: 10.1080/03610730903175782

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283-302. doi: 10.1037/a0023956

Rhodes, M. G. (2016). Judgments of learning. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65-80). doi: 10.1093/oxfordhb/9780199336746.013.4

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review, 27,* 635-643. doi: 10.1007/s10648-015-9332-4

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51,* 71-79. doi: 10.1016/j.jml.2004.03.004

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134,* 124-128. doi: 10.1037/0096-3445.134.1.124

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory & Cognition, 42,* 1260-1272. doi: 10.3758/s13421-014-0431-0

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1258-1266. doi: 10.1037/0278-7393.31.6.1258

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *The Quarterly Journal of*

*Experimental Psychology*, *65*, 2231-2257. doi: 10.1080/17470218.2012.685081

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366. doi: 10.1177/0956797611417632

Son, L. K., & Kornell, N. (2009). Simultaneous decisions at study: Time allocation, ordering, and spacing. *Metacognition and Learning, 4*, 237-248. doi: 10.1007/s11409-009-9049-1

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review, 24*, 379-399. doi: 10.1007/s10648-012-9206-y

Thomas, R. C., Finn, B., & Jacoby, L. L. (2016). Prior experience shapes metacognitive judgments at the category level: The role of testing and category difficulty. *Metacognition and Learning, 11*, 257-274. doi: 10.1007/s11409-015-9144-4

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131. doi: 10.1126/science.185.4157.1124

Van Overschelde, J. P., & Nelson, T. O. (2006). Delayed judgments of learning cause both a decrease in absolute accuracy (calibration) and an increase in relative accuracy (resolution). *Memory & cognition, 34*, 1527-1538. doi: 10.3758/BF03195916

Vaughn, K. E., Hausman, H., & Kornell, N. (2017). Retrieval attempts enhance learning regardless of time spent trying to retrieve. *Memory*, *25*, 298-316. doi: 10.1080/09658211.2016.1170152

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science*, *22*, 1127-1131. doi: 10.1177/0956797611417724

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*, 3-14. doi: 10.1007/s11409-006-6893-0

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*, 568-579. doi: 10.1080/09658211.2012.687052

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611-617. doi: 10.1037/0033-2909.110.3.611

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence, not subjective probability. *Memory & Cognition*, *43*, 1168-1179. doi: 10.3758/s13421-015-0532-4

Zawadzka, K., & Higham, P. A. (2016). Recalibration effects in judgments of learning: A signal detection analysis. *Journal of Memory and Language*, *90*, 161-176. doi: 10.1016/j.jml.2016.04.005

# Appendix A

Ethics Approval Letter

*Sent via email*

Dear Dr Palmer

Ethics Ref No: H0012660
Project title: Confidence in memory

This email is to confirm that the following amendment was approved by the Chair of the Tasmania Social Sciences Human Research Ethics Committee on 15/3/2017:
Addition of student researchers Mr Rod Garton, Ms Amelia Kohl, Ms Morgan Norris, Mr Rafal Kozlowski, Ms Talira Kucina and Ms Rachel Breen.
Removal of student researchers Ms Rebecca Healy, Ms Kate Edwards, Ms Catherine Bishop, Ms Katie-Lee Crawford, Ms Katie Henderson, Ms Rebecca Kaiser, Mr Robert Kirkis, Mr Michael O'Leary and Mr Tane Thomas.

All committees operating under the Human Research Ethics Committee (Tasmania) Network are registered and required to comply with the National Statement on Ethical Conduct in Human Research (NHMRC 2007, updated May 2015).

This email constitutes official approval. If your circumstances require a formal letter of amendment approval, please let us know.

Should you have any queries please do not hesitate to contact me.

Kind regards
Katherine

**Katherine Shaw**
Executive Officer, Social Sciences HREC
Office of Research Services | Research Division
University of Tasmania
Private Bag 1
Hobart TAS 7001
T +61 3 6226 2763
[www.utas.edu.au/research]www.utas.edu.au/research

UNIVERSITY of TASMANIA

# Appendix B

## Information Sheet

Locked Bag 1342 Launceston
Tasmania 7250 Australia
Phone (03) 6324 3004 Fax (03) 6324 3168
matthew.palmer@utas.edu.au

UTAS

## Metacognition for Well-Learned Information

Information Sheet for Participants

**1. Invitation**

You are invited to participate in a psychology experiment examining study strategies and metacognition. The study is being conducted in partial fulfillment of an Honours degree for Talira Kucina under the supervision of Dr Matthew Palmer, within the School of Psychology at the University of Tasmania.

**2. What is the purpose of this study?**

The experiment aims to investigate how different study techniques are related to metacognition, or knowledge about what you think you do and do not know.

**3. Why have I been invited to participate?**

You have been identified on the basis of being 18 years of age or older and a current student at the University of Tasmania or as a member of the wider community.

Participation in this research is completely voluntary meaning you do not have to participate if you do not wish to and there will be no consequences. You are also free to leave the study at any time.

**4. What will I be asked to do?**

On the first occasion, you will be asked to learn lists of Lithuanian-English word pairs and to correctly recall the target English word when the Lithuanian word alone is presented. You will be asked to recall each English word a certain number of times after the list has initially been presented. You will also be asked to provide some ratings on the likelihood that you will be able to recall the English word on a final test about 7 days later. On the second occasion, you will be asked to complete a recall test where you will be shown the Lithuanian cue word and asked to respond with the English word.

The study will take place at the University of Tasmania in a psychology testing room on the Launceston campus and will take approximately 2 hours to complete on the first occasion and 15-30 minutes on the second occasion.

**5. Are there any possible benefits from participation in this study?**

There may be no direct benefits to yourself, however the information gained will provide key insight into psychological theories regarding learning strategies and how people make judgments about what they do and do not know.

For their time, participants will receive either 2.5 hours research credit or $40.00 or a combination thereof.

**6. Are there any possible risks from participation in this study?**

There are no foreseeable risks associated with participating in this research.

Locked Bag 1342 Launceston
Tasmania 7250 Australia
Phone (03) 6324 3004 Fax (03) 6324 3168
matthew.palmer@utas.edu.au

UTAS



**7. What if I change my mind during or after the study?**
You are free to leave the study at any time without giving an explanation. However, once you have started the study it is not possible to withdraw as your information will be stored anonymously and therefore we cannot identify your particular responses.

**8. What will happen to the information when this study is over?**
Data will be securely stored on password protected files on password protected computers for at least five years following the initial reporting of the findings. Following this all data will be archived. Only those conducting the study will have access to the raw data collected.

All data will be stored anonymously in a confidential manner, with no identifying information attached to it.

**9. How will the results of the study be published?**
The research findings will be reported in a Psychology Honours thesis and academic journal.

If you would like to access the final results please contact one of the researchers.

No individual participants will be identified in the publication of this study.

**10. What if I have questions about this study?**
Should you have questions relating to any aspect of this research please feel free to contact Talira Kucina via email: tmkucina@utas.edu.au, or Dr Matthew Palmer via email: matthew.palmer@utas.edu.au.

This study has been approved by the Tasmanian Social Sciences Human Research Ethics Committee. If you have concerns or complaints about the conduct of this study, please contact the Executive Officer of the HREC (Tasmania) Network on +61 3 6226 6254 or email human.ethics@utas.edu.au. The Executive Officer is the person nominated to receive complaints from research participants. Please quote ethics reference number **H0012660**.

**This information sheet is for you to keep. If you wish to participate in this research ask the researcher for a consent form to complete.**

# Appendix C

## Consent Form

Locked Bag 1342 Launceston
Tasmania 7250 Australia
Phone (03) 6324 3004 Fax (03) 6324 3168
matthew.palmer@utas.edu.au

UTAS

UNIVERSITY OF TASMANIA

## Metacognition for Well-Learned Information

*Participant Consent Form*

1. I agree to take part in the research study named above.
2. I have read and understood the Information Sheet for this study.
3. The nature and possible effects of the study have been explained to me.
4. I understand that the first session of the study involves learning lists of word pairs and predicting whether I will be able to later recall them, and that the second occasion involves completing a recall test of the word pairs.
5. I understand that participation involves no foreseeable risks.
6. I understand that all research data will be securely stored on the University of Tasmania premises for five years from the publication of the study results, and will then be destroyed unless I give permission for my data to be archived.

   I agree to have my study data archived. (Note that your data will be stored anonymously.)

   Yes ☐ No ☐
7. Any questions that I have asked have been answered to my satisfaction.
8. I understand that the researchers will maintain confidentiality and that any information I supply to the researcher will be used only for the purposes of the research.
9. I understand that the results of the study will be published so that I cannot be identified as a participant.
10. I understand that my participation is voluntary and that I may withdraw at any time without any effect.

    I understand that I will not be able to withdraw my data after completing the experiment as my data will be anonymous.

Participant's name: _____

Participant's signature: _____

Date: _____

Locked Bag 1342 Launceston
Tasmania 7250 Australia
Phone (03) 6324 3004 Fax (03) 6324 3168
matthew.palmer@utas.edu.au

UTAS

**Statement by Investigator**

☐ I have explained the project and the implications of participation in it to this volunteer and I believe that the consent is informed and that he/she understands the implications of participation.

If the Investigator has not had an opportunity to talk to participants prior to them participating, the following must be ticked.

☐ The participant has received the Information Sheet where my details have been provided so participants have had the opportunity to contact me prior to consenting to participate in this project.

Investigator's name: _____

Investigator's signature: _____

Date: _____

## Appendix D

Transcript of Participant Instructions (Including Demographic and Language Proficiency Questions)

Thank you for taking the time to participate in this experiment.

Please pay full attention throughout and read the instructions carefully. In the study you will be asked to learn several lists of word pairs which you will then be tested on. To begin with there are some demographic questions for you to complete.

Please click continue to proceed.

Demographic Questions

Please type in your response to the following questions:

Age:

Gender:

Level of education:

Is English your first language?

Can you speak Lithuanian?

Can you speak a language other than English? If yes, please specify:

Please click continue to proceed.

*Study phase*

Now you have completed these questions, the experiment will begin.

First you will be asked to learn a list of Lithuanian-English word pairs. Each word pair will be displayed on the screen one at a time.

e.g., KNYGA – BOOK

After all items in this list are presented there will be several test-restudy trials that you will be asked to complete. This will consist of the Lithuanian word being shown on screen and you typing in the English translation.

Please click continue to proceed.

You will now be presented with the first list of Lithuanian-English word pairs. Each item will be displayed on the screen for 6 seconds before automatically moving onto the next item.

Please click continue when you are ready to begin.

[List 1 presented]

You have now completed the first study trial.

Next you will begin the test-restudy period.

Please click continue to proceed.

*Practice test – restudy phase*

A Lithuanian word will appear on the screen and you will have 8 seconds to type in the corresponding English word before automatically moving onto the next item.

e.g., KNYGA - [ ]

If you do not type a response, or respond incorrectly, the correct word will appear on screen before automatically moving on.

e.g., KNYGA - [BOOK]

Please take care when spelling the English word to make sure it is as accurate as possible.

Please click continue to proceed.

Word pairs will be repeated until an acceptable level of performance has been reached.

Immediately after an item has be recalled a sufficient number of times, you will be asked to predict how well you will remember the information.

You will be prompted to choose a response regarding the likelihood that you will remember that word pair in about 7 days. Here, 0 means you are not at all likely to remember the English word and 100 means you are definitely likely to remember the English word. A score of 50 means you are just as likely to remember the English word as you are not to remember it. (Modified for binary condition: select NO if you think you are not likely to remember the English word, or YES if you think you are likely to remember the English word.)

Please click continue when you are ready to begin.

[List 1 presented for restudy]

For the item you just saw, how likely are you (changed to are you likely for binary condition) to correctly recall the English word when presented with the Lithuanian word alone on a test in about 7 days?

Please select a response from 0 (0% likelihood of recalling item) to 100 (100% likelihood of recalling item).

*OR*

Please select a response of YES (I think I will be able to recall the item) or NO (I do not think I will be able to recall the item).

[Select response from 0-100 *OR* yes/no]

You have now completed the first test-restudy phase.

Next you will be shown another list of word pairs for an initial study period, just as you did before. Again, these will be displayed for 6 seconds each before automatically moving on.

Once all items have been presented, you will then be tested on them.

Please click continue when you are ready to begin.

[Study phase and practice test – restudy phase repeated for remaining lists]

*Halfway through the experiment, irrespective of lag and list order*

You have now completed the first half of the experiment.

There will now be a 2 minute break for you to rest or stretch your legs.

*Final screen displayed*

You have now completed the testing session for today.

Please let the experimenter know when you reach this point.

*Cued-recall task* (approximately 7 days later)

Welcome to the second session of the experiment.

Today you will be shown all the Lithuanian words that were presented last time and asked to respond with the corresponding English word.

Please click next to continue.

The Lithuanian words will be displayed on the screen one at a time and there will be a space provided for you to type the English word. There is no time limit here. All words will be presented in one group, irrespective of the lists in which you initially learned them.

Please make sure your spelling is as accurate as possible.

Once you have provided a response, click Next to move onto the next word. If you do not know the answer, type "X", and then click Next to move onto the next word. Once you have progressed past a word you will not have a chance to return to it. Please click Next to begin.

[Presentation of word list]

You have now completed the final testing session.

Thank you for your participation!

Please let the experimenter know when you reach this point.

## Appendix E

Lithuanian-English Word Pairs

| Lithuanian | English | Lithuanian | English |
|---|---|---|---|
| Lova | Bed | Tiltas | Bridge |
| Rūsys | Basement | pliažas | Beach |
| Tinklas | Net | Traukinys | Train |
| Upė | River | Sesuo | Sister |
| Karalius | King | Pupa | Bean |
| Sausainis | Cookie | Palaidinė | Shirt |
| Namas | House | Daina | Song |
| Želė | Jelly | Akis | Eye |
| Nafta | Oil | Smegenys | Brain |
| Pomidoras | Tomato | Mėsa | Meat |
| Burna | Mouth | Gėlė | Flower |
| Mokykla | School | Pastatas | Building |
| Riteris | Knight | Auksas | Gold |
| Padanga | Tyre | Būgnas | Drum |
| Paukštis | Bird | Arbata | Tea |
| Žolė | Grass | Vanduo | Water |
| Lietus | Rain | Vilkas | Wolf |
| Langas | Window | Koja | Leg |
| Urvas | Cave | Vinis | Nail |
| Augalas | Plant | Puodelis | Cup |
| Adata | Needle | Kumpis | Ham |
| Mėnulis | Moon | Duona | Bread |
| Bulvė | Potato | Žiedas | Ring |
| Medus | Honey | Piniginė | Wallet |
| Muilas | Soap | Kriauklė | Sink |
| Laikrodis | Clock | Kardas | Sword |
| Vonia | Bath | Geležis | Iron |
| Krauias | Blood | Šepetys | Brush |
| Laidas | Wire | Kirvis | Axe |
| Krosnis | Stove | Maišas | Bag |
| Vėliava | Flag | Padažas | Gravy |
| Diržas | Belt | Kablelis | Hook |
| Smuikas | Violin | Šalmas | Helmet |
| Krantas | Shore | Kamuolys | Ball |
| Stalas | Table | Laiškas | Letter |
| Voras | Spider | Varpas | Bell |

# Appendix F

Results Tables

Table F1

*Main Effect and Interactions for Recall Accuracy*

| Variables | *F* | df | *p* | *f* |
|---|---|---|---|---|
| Scale | 0.11 | 1, 44 | .739 | .049 |
| Criterion × Scale | 0.08 | 2, 88 | .928 | .016 |
| Lag × Scale | 2.21 | 1, 44 | .144 | .091 |
| Criterion × Lag × Scale | 0.17 | 2, 88 | .841 | .007 |

Table F2

*Main Effect and Interactions for Metacognitive Judgments (JOLs)*

| Variables | *F* | df | *p* | *f* |
|---|---|---|---|---|
| Scale | 0.00 | 1, 44 | .990 | .001 |
| Lag × Scale | 0.25 | 1, 44 | .619 | .029 |
| Criterion × Scale | 2.06 | 2, 88 | .145 | .073 |
| Criterion × Lag | 0.85 | 2, 88 | .432 | .034 |
| Criterion × Lag × Scale | 0.05 | 2, 88 | .948 | .002 |

Table F3

*Interactions for Absolute Accuracy*

| Variables | *F* | df | *p* | *f* |
|---|---|---|---|---|
| Measure × Scale | 0.03 | 1, 44 | .856 | .018 |
| Criterion × Measure × Scale | 1.75 | 2, 88 | .180 | .045 |
| Lag × Measure × Scale | 1.90 | 1, 44 | .175 | .062 |
| Criterion × Lag × Measure × Scale | 0.07 | 2, 88 | .929 | .021 |

Table F4

*Descriptive Statistics for each Lag and Criterion Combination for JOLs and Recall Accuracy*

| | JOLs | | Accuracy | |
|---|---|---|---|---|
| Criterion | *M* (*SD*) | 95% CI | *M* (*SD*) | 95% CI |
| Long lag | | | | |
| 1 | 27.74 (26.72) | [19.80, 35.67] | 19.38 (16.95) | [14.35, 24.42] |
| 3 | 41.23 (32.65) | [31.54, 50.93] | 35.69 (22.61) | [28.97, 42.40] |
| 9 | 58.48 (31.98) | [48.98, 67.97] | 50.18 (26.03) | [42.45, 57.91] |
| Short lag | | | | |
| 1 | 34.67 (27.68) | [26.45, 42.89] | 10.51 (12.35) | [6.84, 14.17] |
| 3 | 48.77 (31.15) | [39.52, 58.02] | 15.58 (18.05) | [10.22, 20.94] |
| 9 | 61.50 (32.51) | [51.85, 71.16] | 20.83 (18.82) | [15.24, 26.42] |

Table F5

*Main Effect and Interactions for Relative Accuracy*

| Variables | *F* | df | *p* | *f* |
|---|---|---|---|---|
| Criterion | 1.04 | 2, 88 | .356 | .121 |
| Criterion × Scale | 1.97 | 2, 88 | .146 | .226 |
| Lag × Scale | 2.63 | 1, 44 | .112 | .168 |
| Criterion × Lag | 0.19 | 2, 88 | .825 | .054 |
| Criterion × Lag × Scale | 0.35 | 2, 88 | .706 | < .001 |